D76–D81 *Nucleic Acids Research, 2015, Vol. 43, Database issue* *Published online 27 September 2014*
doi: 10.1093/nar/gku887

AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors

Hong-Mei Zhang†, Teng Liu†, Chun-Jie Liu, Shuangyang Song, Xiantong Zhang, Wei Liu, Haibo Jia, Yu Xue and An-Yuan Guo*

Department of Biomedical Engineering, Key Laboratory of Molecular Biophysics of the Ministry of Education, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, PR China

Received August 17, 2014; Revised September 12, 2014; Accepted September 12, 2014

ABSTRACT

Transcription factors (TFs) are key regulators for gene expression. Here we updated the animal TF database AnimalTFDB to version 2.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/). Using the improved prediction pipeline, we identified 72 336 TF genes, 21 053 transcription co-factor genes and 6502 chromatin remodeling factor genes from 65 species covering main animal lineages. Besides the abundant annotations (basic information, gene model, protein functional domain, gene ontology, pathway, protein interaction, ortholog and paralog, etc.) in the previous version, we made several new features and functions in the updated version. These new features are: (i) gene expression from RNA-Seq for nine model species, (ii) gene phenotype information, (iii) multiple sequence alignment of TF DNA-binding domains, and the weblogo and phylogenetic tree based on the alignment, (iv) a TF prediction server to identify new TFs from input sequences and (v) a BLAST server to search against TFs in AnimalTFDB. A new nice web interface was designed for AnimalTFDB 2.0 allowing users to browse and search all data in the database. We aim to maintain the AnimalTFDB as a solid resource for TF identification and studies of transcription regulation and comparative genomics.

INTRODUCTION

Transcription factors (TFs) are key regulators of gene expression in all organisms. They are usually classified into different families by their DNA-binding domains (DBDs). Usually, there are more than 5% TF genes in vertebrates and angiosperms (1,2). It is estimated that human genome contains ∼1700 TF genes, occupying more than 7% of the protein-coding genes (3). Similar as the studies of plant TF databases (4–6), there are several databases for TFs in one or more animal genomes, such as Riken mouse TFdb (7), FlyTF (8), TFCat (9), TFCONES (10), ITFP (11) and DBD (12). However, all these databases were built before 2010 and were not updated in recent years. In 2011, we characterized the TF families and constructed a comprehensive animal TF database (AnimalTFDB) (2), which contains TFs, co-factors and chromatin remodeling factors (CRFs) in 50 animal species. The AnimalTFDB database was accessed thousands of times and widely used for functional and evolutionary studies.

Recent advance in high-throughput transcriptome sequencing (RNA-Seq) provides powerful ways to quantify the gene expression in a sample. There are many expression data sequenced for different tissues of human and model species, such as the human body map project (13), TCGA project (14) and other studies about the evolution of gene expression (15,16). Thus, it is feasible and very useful to explore the expressions of TFs from these RNA-Seq data. In the past 3 years, many genomes were sequenced and the number of species in Ensembl database was increased by more than a quarter (17). Thus, an updated animal TF database including the data of newly sequenced genomes is needed and an online animal TF prediction server is very necessary.

To meet the data-driven research requirements, we improved the prediction pipeline and updated AnimalTFDB to version 2.0 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/). In comparison with the previous version, AnimalTFDB 2.0 covers more species and new types of annotations including gene phenotype and expression data in nine species. An online TF prediction server was set up. The multiple sequence alignment of TF DBD sequences and phylogenetic trees for each TF family of every species were also constructed. Taken together, AnimalTFDB 2.0 provides users with comprehensive animal TF lists, annotations and prediction tools.

*To whom correspondence should be addressed. Tel: +86 27 8779 3177; Fax: +86 27 8779 3177; Email: guoay@mail.hust.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

MATERIALS AND METHODS

Data sources

We downloaded all the protein sequences of 65 animal genomes from Ensembl (version 75) (17) to identify their TFs, transcription co-factors and CRFs. We obtained most of the gene annotations from NCBI Entrez Gene and Ensembl databases, which includes basic information, orthologs, paralogs, phenotype, Gene Ontology (GO) and gene model. The protein–protein interaction information was parsed from BioGRID (18) and HPRD (19) databases. The pathway annotations were extracted from BioCarta (http://www.biocarta.com/) and KEGG databases. Putative functional domains were searched by PfamScan (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/) program.

Rich information for gene expression is provided in AnimalTFDB 2.0. We downloaded the human gene expression data of cancers, tissues and cell lines from TCGA (https://tcga-data.nci.nih.gov/tcga/findArchives.htm) and EBI Expression Atlas (http://www.ebi.ac.uk/gxa/download.html). The expression data of the human proteome were parsed from two recent Nature papers (20,21). The gene expression of *Drosophila melanogaster* and *Caenorhabditis elegans* was extracted from the data published by Li *et al.* (22). Our collaborators Drs Yu Xue and Haibo Jia kindly provided the unpublished gene expression data of Danio rerio. We downloaded the raw data for Rattus norvegicus, Bos taurus and Gallus gallus from NCBI GEO DataSets published by Burge group (16) and estimated gene expression levels with TopHat (23) and Cufflinks (24) programs. The gene expression data for Mus musculus and Macaca mulatta were downloaded from RhesusBase (25,26), which were estimated from the RNA-Seq data published by groups Burge (16), Kaessmann (15) and Chuan-Yun Li (25).

Animal TF family and assignment rules

TFs are usually characterized and classified into specific families by their DBDs. After reviewing recently published literature, we found two new TF families NCU-G1 and CEP-1 comparing with AnimalTFDB 1.0, while CEP-1 only exists in *C. elegans*. In addition, the nuclear receptor superfamily was reclassified, which was grouped into 12 subfamilies based on InterPro (27) and Pfam (28) annotations in AnimalTFDB 1.0. In the updated version, we classified it into seven subfamilies according to the classification method of nuclear receptor nomenclature committee (29). The nuclear TF Y (NFY) was also classified into three subfamilies based on its three different subunits. In AnimalTFDB 2.0, there are 70 TF families and one of them named 'Others' including some orphan TFs.

In most cases, a TF only has one type of DBD, thus it is easy to assign it into a certain family correctly. But in some cases, a TF may have more than one type of DBD. In order to classify them into correct family, we checked all the TFs of human and mouse which have multiple types of DBDs, and then set up two rules. First, if a superfamily contains several subfamilies, we classified the TFs based on the subfamily DBD. For example, the Homeobox superfamily has four subfamilies: Pou, CUT, TF_Otx and other Homeobox. In this superfamily, all TFs have a Homeobox domain, and some of them have one of the Pou, CUT and TF_Otx subfamily signature domains. We assigned them into specific family based on their subfamily signature domain. The second rule is that if a TF has more than one unrelated DBD, we will classify it into the family based on the DBD with the smallest E-value. We checked the classification of human and mouse TFs and found our method was reasonable.

TF prediction pipeline

We refined the TF prediction pipeline by updating the hidden Markov model (HMM) profiles of TF DBDs and adjusting the TF family assignment rules. The latest HMM profiles for most DBDs were downloaded from Pfam version 27.0 (28). For the remaining domains without available Pfam HMM profiles, we rebuilt the HMM profiles using the sequences in representative species (human, mouse, zebrafish and fly). To predict TFs, we applied the hmmsearch program in HMMER 3.0 (30) to search all the protein sequences in each species against the HMM profiles with E-value 0.0001 as the cutoff. Then we assigned the TFs into different families according to the above family assignment rules.

Identification of transcription co-factors and CRFs

We also adjusted the identification method of transcription co-factors and CRFs. First, we extracted both of them for human from Tcof-DB (31) and GO database by related GO items. For transcription co-factors, the used GO items are 'transcription coactivator activity', 'transcription corepressor activity', 'transcription co-factor activity' and 'regulation of transcription'. For CRFs, the GO annotations are 'chromatin remodeling', 'chromatin-mediated maintenance of transcription', 'histone *ylation', 'histone .*ylase activity' and 'histone *transferase activity'. After manual curation and removing redundant genes, 415 transcription co-factors and 142 CRFs were obtained in human genome. To identify them in other 64 species, we did the reciprocal best-hits Basic Local Alignment Search Tool (BLAST) between human and other species with the threshold setting as E-value $\leq$ 1e-4, coverage $\geq$ 50% and identity $\geq$ 30%.

RESULTS

Genomic repertoires of three kinds of regulatory factors

Using the refined prediction pipeline, we identified 72 336 TFs, 21 053 transcription co-factors and 6502 CRFs in 65 animal species (Table 1). Their numbers and percentages in model species are shown in Table 2. As a result, almost all of the vertebrates have 5–8.9% of TF genes in their genomes and the proportion of TFs in invertebrates is less than 5% (Supplementary Table S1). The large increase of TF percentage in vertebrates compared to invertebrates is due to the two-rounds of whole genome duplication that occurred in the stem vertebrate lineage followed by retention of a higher number of TF duplicates (32,33). Among the vertebrates, the zebrafish has the most TF genes (2345) and TF percentage (8.86%), because it retained more TF genes after the additional whole genome duplication (3R) in the teleost ancestor (34,35). In addition, the percentages of transcription co-factors and CRFs in vertebrates are about 1.8% and

Table 1. Comparison of data contents between two versions of AnimalTFDB

AnimalTFDB	Version 1.0	Version 2.0
Species	50	65
TF families	72	70
TF genes	52 722	72 336
Co-factor genes	9066	21 053
CRFs genes	3476	6502
Annotation		
-gene function description	No	Yes
-expression	No	Yes
-phenotype	No	Yes
Multi-alignment of DBDs and their WebLogo	No	Yes
Phylogenetic tree	No	Yes
TF prediction server	No	Yes
BLAST search	No	Yes

Table 2. Summary of the expression data and TF numbers of model species in AnimalTFDB 2.0

Species	Lineage	Expression[a]	TF (%)[b]	Expressed TF (%)[c]	Co-factor (%)[b]	Expressed co-factor (%)[c]	CRF (%)[b]	Expressed CRF (%)[c]
Homo sapiens	Primate	CA (27), TI (16,24), CL (22)	1691 (7.4%)	1589 (94.0%)	462 (2.0%)	430 (93.1%)	155 (0.7%)	140 (90.3%)
Macaca mulatta	Primate	TI (11)	1418 (6.5%)	964 (68.0%)	378 (1.7%)	291 (77.0%)	118 (0.5%)	95 (80.5%)
Mus musculus	Rodent	TI (10)	1485 (6.5%)	1227 (82.6%)	397 (1.7%)	390 (98.2%)	122 (0.5%)	118 (96.7%)
Rattus norvegicus	Rodent	TI (9)	1375 (6.0%)	1137 (82.7%)	382 (1.7%)	374 (97.9%)	118 (0.5%)	116 (98.3%)
Bos taurus	Laurasiatheria	TI (9)	1280 (6.4%)	1141 (89.1%)	378 (1.9%)	376 (99.5%)	121 (0.6%)	121 (100.0%)
Gallus gallus	Bird	TI (9)	858 (5.5%)	769 (89.6%)	329 (2.1%)	325 (98.8%)	98 (0.6%)	98 (100.0%)
Danio rerio	Fish	DS (8)	2345 (8.9%)	1756 (74.9%)	315 (1.2%)	306 (97.1%)	100 (0.4%)	97 (97.0%)
Drosophila melanogaster	Insect	TI (29), CL (19), DS (30)	604 (4.3%)	594 (98.3%)	160 (1.1%)	158 (98.8%)	53 (0.4%)	51 (96.2%)
Caenorhabditis elegans	Nematoda	TI (4), CT (14), DS (35)	706 (3.4%)	684 (96.9%)	130 (0.6%)	130 (100.0%)	40 (0.2%)	39 (97.5%)

[a]CA, cancer; TI, tissue; DS, development stage; CL, cell line; CT, cell type. Number in the bracket is the number of data sets of that type. The TI (16,24) of human indicates there are 16 mRNA data sets and 24 protein data sets for human tissue expression data. All other gene expression data are from RNA-seq mRNA expression.
[b]The percentages in brackets are the percentages of TF (co-factor or CRF) genes in the protein-coding genes of genomes.
[c]The percentages in brackets are the percentages of expressed TF (co-factor or CRF) genes.

0.6% of their protein-coding genes on average, which are also higher than those of invertebrates.

Comprehensive annotations

In an attempt to construct a comprehensive knowledgebase for animal TFs, we provided rich information for them. Besides, the abundant annotations provided in version 1.0, we collected gene function description, gene expression at mRNA and protein levels, and phenotype data from various public resources and performed annotation for these factors (Figure 1). Through checking the transcription regulation-related GO annotation with experimental evidence codes, we marked the regulators as experimentally validated or putative in seven model species. As a result, we found 426 TFs, 236 co-factors and 37 CRFs with experimental evidence in human. In addition, using the DBD sequences, we made multiple sequence alignment by ClustalW2 (36) and constructed phylogenetic trees for TFs in each family of each species by applying neighbor-joining method in PHYLIP package (37) with bootstrap 100. The multiple sequence alignment result and phylogenetic tree were displayed by Weblogo (38) and Phylogeny.fr (39), respectively (Figure 1A). The phylogenetic tree will be helpful for users to infer the functions of poorly studied TFs.

Gene expression

The gene expression information of nine species is provided in AnimalTFDB 2.0 involving normal tissues, cell lines, development stages and cancers in human (Table 2, Figure 1D). We considered a gene is expressed with RPKM $\geq$ 0.5 according to the benchmark set by Xie *et al.* (40). More than 90% of the co-factors and CRFs were detected to be expressed in at least one sample except for *Macaca mulatta*, which may be caused by its different gene annotation between Ensembl and UCSC. However, the percentage of expressed TFs is lower compared with co-factors and CRFs in most of the species. We also made a general analysis for the TF expression pattern in 16 human normal tissues. More than 50% of TFs are expressed in at least 14 tissues and 32% of TFs are expressed in all 16 tissues, such as YBX1, YBX3, EGR1, ATF4, FOS, JUN and MYC. More than 7% of TFs are only expressed in one tissue and most of them are expressed at low levels. The numbers of expressed TFs are also different between tissues, ranging from 800 in liver to 1200 in testis.

TF prediction server

With the development of high-throughput sequencing technology, a growing number of genomes and transcriptomes are being or will be sequenced. A TF prediction server will be helpful for users to identify TFs from their own protein sequences. In this regard, we set up a TF prediction server (http://bioinfo.life.hust.edu.cn/AnimalTFDB/prediction.shtml) in AnimalTFDB 2.0 (Figure 1B). Same prediction method and TF family assignment rules described above were used for this server. In the prediction result page, TF family, alignment e-value and detailed align-

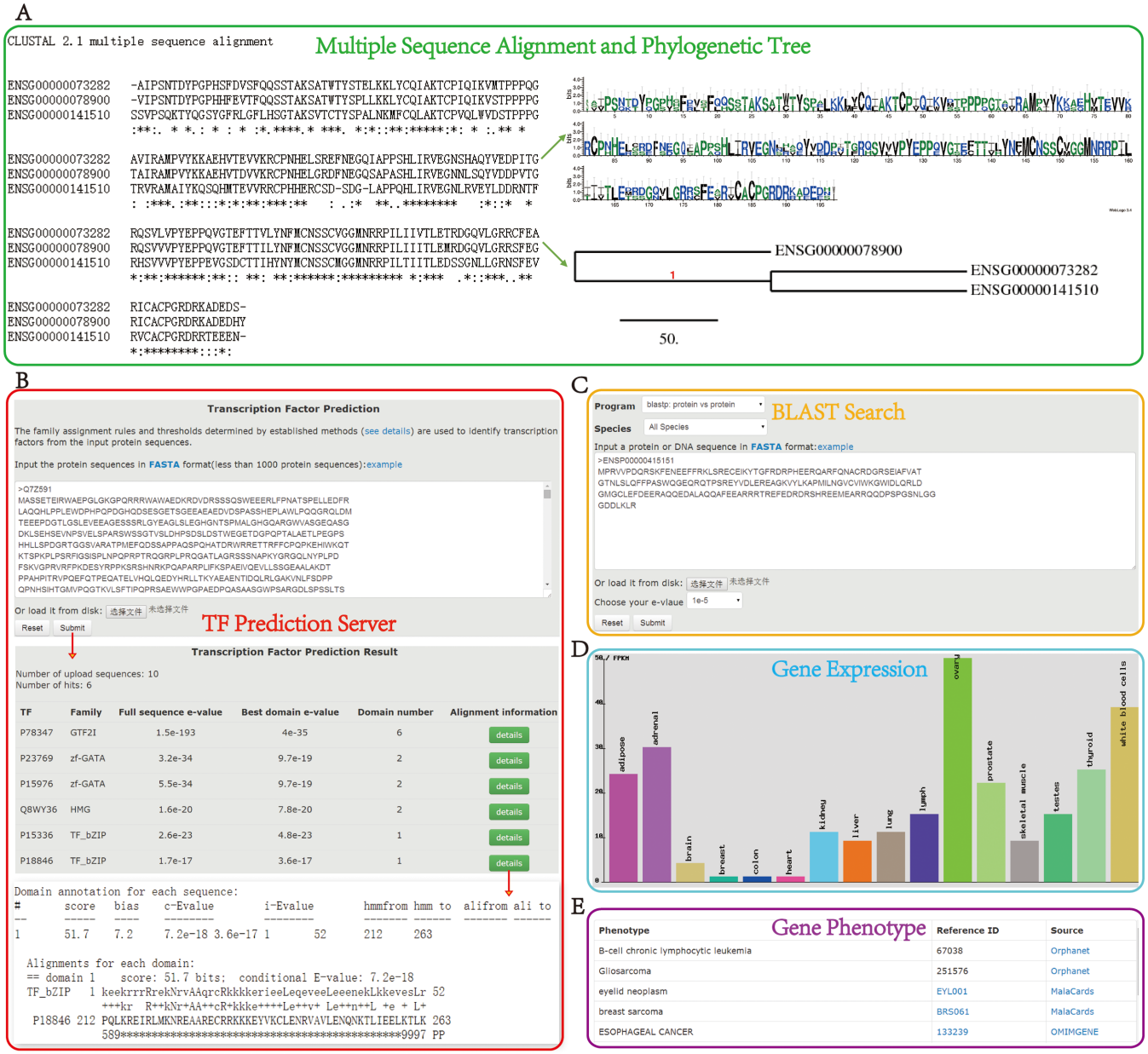

Figure 1. The new annotations and tools in AnimalTFDB 2.0. (A) The multiple sequence alignment of TF DBDs, the weblogo and phylogenetic tree based on the alignment in each TF family. (B) The TF prediction server and examples of prediction result. (C) The BLAST search server. (D) One example of gene expression information. (E) The gene phenotype information.

ment information will be provided. Currently, users can upload up to 1000 protein sequences and obtain results within a few minutes from our server.

BLAST search

To help users find homologous gene and explore functions of poorly studied TFs, we provided a BLAST tool (http://bioinfo.life.hust.edu.cn/AnimalTFDB/blast.shtml) to search against TFs in our database with protein or DNA sequences (Figure 1C). The protein sequences of all species or one specific species could be selected as the BLAST database.

SUMMARY AND FUTURE PERSPECTIVES

We have updated our AnimalTFDB to version 2.0, which provides TF, transcription co-factor and CRF repertoires in 65 species across 11 lineages. The abundant annotation, gene expression profiles and phylogenetic trees will be useful resources for further exploration of the physiological function and evolutionary relationship of TFs. In addition, the TF prediction server in AnimalTFDB 2.0 will be helpful for TF identification in newly sequenced genome. In the future, we will continue to work on this project in the following directions: refining the TF family assignment rules and prediction pipeline, collecting more types of useful annotations for identified regulators, adding more species when new an-

imal genome data is available and keeping the web interface compact, clear and beautiful. We aim to maintain a comprehensive animal TF database for a long time to provide a solid resource for the studies of transcriptional regulation and comparative genomics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Drs Haiyan Huang and Jingyi Jessica Li for offering gene expression data of *D. melanogaster* and *C. elegans* to us and thank Prof. Chuan-Yun Li for providing the gene expression data of *Macaca mulatta* and *Mus musculus*. We are also grateful to our users and all members in our lab for their valuable suggestions and comments.

FUNDING

National Natural Science Foundation of China (NSFC) [31171271, 31270885, 31471247]; Program for New Century Excellent Talents in University (NCET), Ministry of Education of China. Funding for open access charge: National Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Jin,J., Zhang,H., Kong,L., Gao,G. and Luo,J. (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.*, **42**, D1182–D1187.
2. Zhang,H.M., Chen,H., Liu,W., Liu,H., Gong,J., Wang,H. and Guo,A.Y. (2012) AnimalTFDB: a comprehensive animal transcription factor database. *Nucleic Acids Res.*, **40**, D144–D149.
3. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
4. Guo,A.Y., Chen,X., Gao,G., Zhang,H., Zhu,Q.H., Liu,X.C., Zhong,Y.F., Gu,X., He,K. and Luo,J. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, **36**, D966–D969.
5. Charoensawan,V., Wilson,D. and Teichmann,S.A. (2010) Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Res.*, **38**, 7364–7377.
6. Guo,A., He,K., Liu,D., Bai,S., Gu,X., Wei,L. and Luo,J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.
7. Kanamori,M., Konno,H., Osato,N., Kawai,J., Hayashizaki,Y. and Suzuki,H. (2004) A genome-wide and nonredundant mouse transcription factor database. *Biochem. Biophys. Res. Commun.*, **322**, 787–793.
8. Pfreundt,U., James,D.P., Tweedie,S., Wilson,D., Teichmann,S.A. and Adryan,B. (2010) FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res.*, **38**, D443–D447.
9. Fulton,D.L., Sundararajan,S., Badis,G., Hughes,T.R., Wasserman,W.W., Roach,J.C. and Sladek,R. (2009) TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol.*, **10**, R29.
10. Lee,A.P., Yang,Y., Brenner,S. and Venkatesh,B. (2007) TFCONES: a database of vertebrate transcription factor-encoding genes and their associated conserved noncoding elements. *BMC Genom.*, **8**, 441.
11. Zheng,G., Tu,K., Yang,Q., Xiong,Y., Wei,C., Xie,L., Zhu,Y. and Li,Y. (2008) ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics*, **24**, 2416–2417.
12. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.
13. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
14. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
15. Brawand,D., Soumillon,M., Necsulea,A., Julien,P., Csardi,G., Harrigan,P., Weier,M., Liechti,A., Aximu-Petri,A., Kircher,M. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
16. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**, 1593–1599.
17. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
18. Chatr-Aryamontri,A., Breitkreutz,B.J., Heinicke,S., Boucher,L., Winter,A., Stark,C., Nixon,J., Ramage,L., Kolas,N., O'Donnell,L. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
19. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
20. Kim,M.S., Pinto,S.M., Getnet,D., Nirujogi,R.S., Manda,S.S., Chaerkady,R., Madugundu,A.K., Kelkar,D.S., Isserlin,R., Jain,S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.
21. Wilhelm,M., Schlegl,J., Hahne,H., Moghaddas Gholami,A., Lieberenz,M., Savitski,M.M., Ziegler,E., Butzmann,L., Gessulat,S., Marx,H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
22. Li,J.J., Huang,H., Bickel,P.J. and Brenner,S.E. (2014) Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.*, **24**, 1086–1101.
23. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
24. Trapnell,C., Hendrickson,D.G., Sauvageau,M., Goff,L., Rinn,J.L. and Pachter,L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, **31**, 46–53.
25. Zhang,S.J., Liu,C.J., Yu,P., Zhong,X., Chen,J.Y., Yang,X., Peng,J., Yan,S., Wang,C., Zhu,X. *et al.* (2014) Evolutionary interrogation of human biology in well-annotated genomic framework of rhesus macaque. *Mol. Biol. Evol.*, **31**, 1309–1324.
26. Zhang,S.J., Liu,C.J., Shi,M., Kong,L., Chen,J.Y., Zhou,W.Z., Zhu,X., Yu,P., Wang,J., Yang,X. *et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res.*, **41**, D892–D905.
27. McDowall,J. and Hunter,S. (2011) InterPro protein classification. *Methods Mol. Biol.*, **694**, 37–47.
28. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
29. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
30. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
31. Schaefer,U., Schmeier,S. and Bajic,V.B. (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, **39**, D106–D110.
32. Huminiecki,L. and Heldin,C.H. (2010) 2R and remodeling of vertebrate signal transduction engine. *BMC Biol.*, **8**, 146.

33. Conant,G.C. and Wolfe,K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.*, **9**, 938–950.
34. Gillis,W.Q., St John,J., Bowerman,B. and Schneider,S.Q. (2009) Whole genome duplications and expansion of the vertebrate GATA transcription factor gene family. *BMC Evol. Biol.*, **9**, 207.
35. Blomme,T., Vandepoele,K., De Bodt,S., Simillion,C., Maere,S. and Van de Peer,Y. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, **7**, R43.
36. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
37. Lim,A. and Zhang,L. (1999) WebPHYLIP: a web interface to PHYLIP. *Bioinformatics*, **15**, 1068–1069.
38. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
39. Dereeper,A., Guignon,V., Blanc,G., Audic,S., Buffet,S., Chevenet,F., Dufayard,J.F., Guindon,S., Lefort,V., Lescot,M. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.
40. Xie,C., Zhang,Y.E., Chen,J.Y., Liu,C.J., Zhou,W.Z., Li,Y., Zhang,M., Zhang,R., Wei,L. and Li,C.Y. (2012) Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.*, **8**, e1002942.